Student:Bingyang Dou Subject:CIS576 Assignment 4:Text Analysis/Visualization and Network Visualization

This is an individual assignment. You are welcome to discuss the assignment with your colleagues, but the work you turn in must be your own.

Find a "corpus" of text that is interesting to you, and perform an analysis and word cloud visualization. There are speech transcripts from both parties that might be interesting!

a. R: Be sure to remove words that should be removed… submit the R code, word clouds, analysis supporting why certain words were removed. (15 points)

First, I decided my topic as the 'The Important Speeches in Steve Jobs' Life'. I want to find out the differences among his important speeches in his legendary life and find out the keys why the speeches are always attractive and successful.

Following this idea, I searched a lot of Steve Jobs' speeches. Finally, I selected these five speeches because they are all remarkable, technically professional, and popular.

- Steve Jobs introduces the Original Macintosh - Apple Shareholder Event (1984-01-24)
- Steve Jobs introduces the Original iMac - Apple Special Event (1998-05-06)
- Steve Jobs introduces the Original iPod - Apple Special Event (2001-10-23)
- Steve Jobs introduces the Original iPhone at Macworld SF (2007-01-09)
- Steve Jobs introduces the Original iPad - Apple Special Event (2010-01-07)

I downloaded the subtitles of those speeches from YouTube and transformed them to txt files. (Using website http://keepvid.com/ to download SRT files.) I also deleted the parts where are not the Steve Jobs' speaking and trimmed a little bit appropriately. Although I could get whatever I want in this way, it also has the shortcomings. The source subtitles are not exact because the website can't recognize mistakes. What's more, there is no space between line and line. For these reasons, I spent six hours on checking and trimming the subtitles.

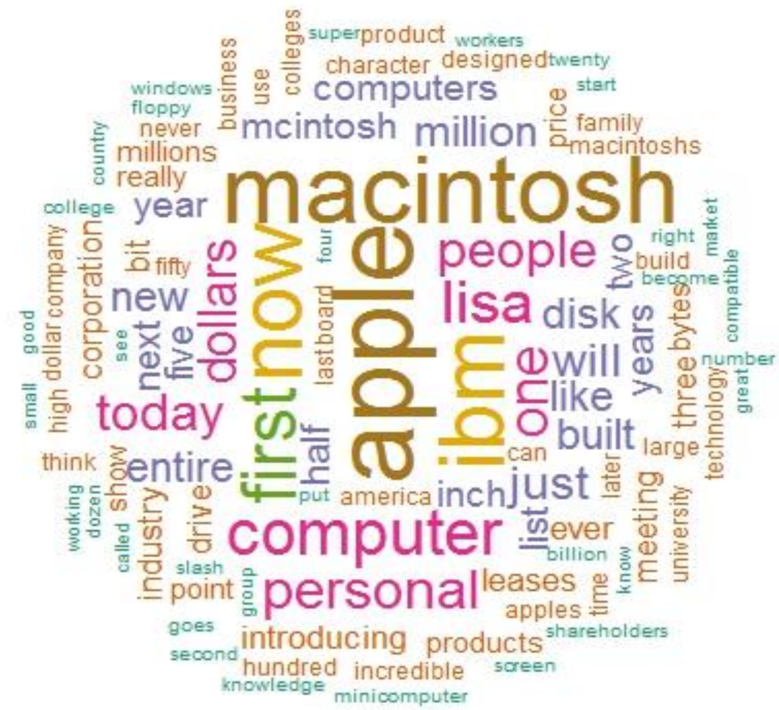
Removing analysis:

Because I generated the word files from the subtitles, there were a lot of timestamps. I removed them by using 'removeNumbers' and 'removePunctuation' because the timestamps were composited by numbers and punctuation. It also removed the numbers and punctuation in all places. Furthermore, I removed the unnecessary whitespace, converted everything to lower case, and removed common words by using 'stripWhitespace', 'tolower', as well as 'removeWords'.

I didn't remove his other words in the first three speeches because I found that almost every word has its special meaning. For example, I had ever considered removing 'thing' and 'things' in the second speech because it looks no attitude. However, I finally found that his second speech was special and the 'thing' and 'things' were special, too. I couldn't remove them. It was because, at that time, iMac was a new product. There was no definition of iMac. That's the way of Steve Jobs' introduction. He wanted to introduce the functions and the features first, leading to the concept of iMac. iMac was only the name of that thing, but that thing could realize a lot of 'was-impossible'. I removed 'just' in the fourth and the fifth speech because I couldn't think of any attitude or meaning.

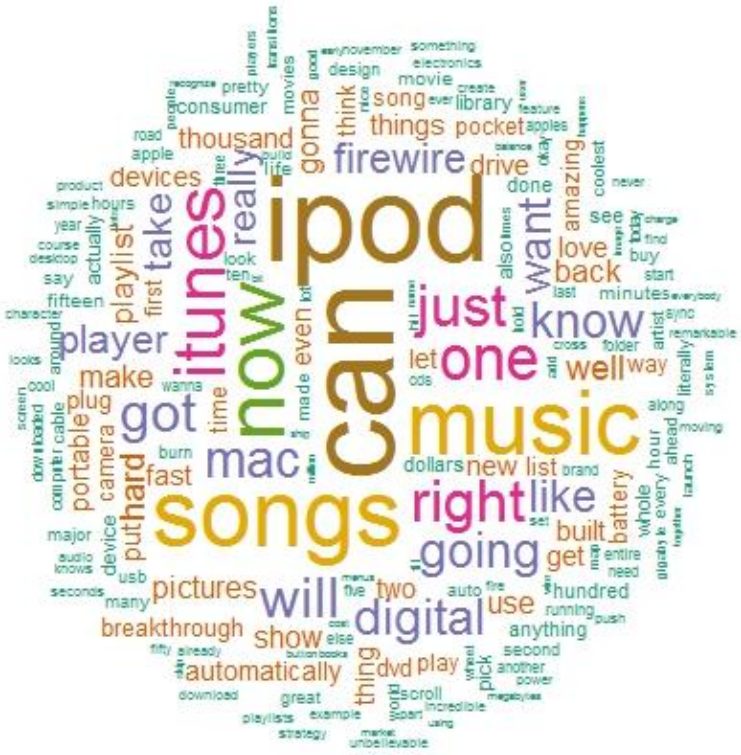
Student:Bingyang Dou Subject:CIS576 Assignment 4:Text Analysis/Visualization and Network Visualization

1. Macintosh

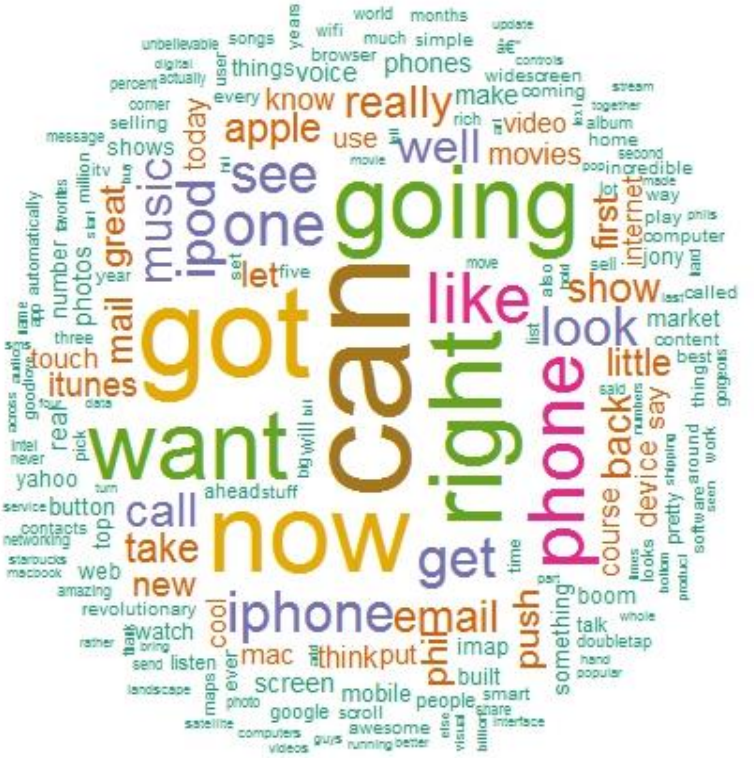


2. iMac

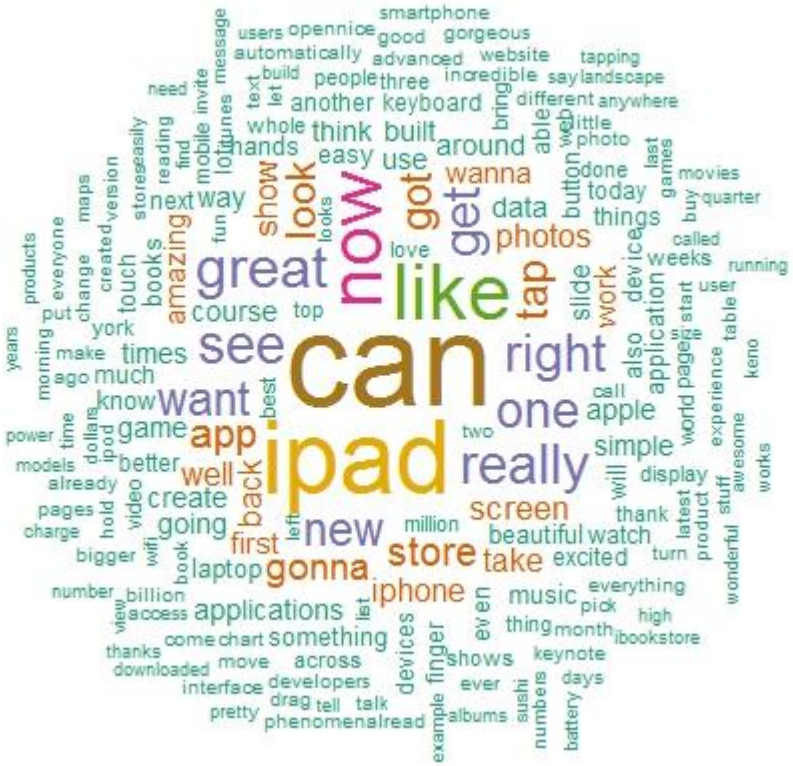
3. iPod



4. iPhone



5. iPad



Analysis I:

To begin with, by using the word cloud, I found that Steve Jobs repeated the new names of those products very frequently, except the second speech. For example, we can see ‘apple’ and ‘Macintosh’ very big in first speech. In the third, fourth, and fifty, we can find ‘iPod’, ‘phone’, ‘iPad’ in very big size. However, in the second speech, I can only find ‘iMac’ in the fourth big class. It maybe because he was trying to introduce the amazing functions of iMac first in the second speech.

Second, in these five speeches, he used three times ‘can’ in first-frequent-class but only twice include ‘apple’ in the first class. Comparing with ‘apple’, he repeated more ‘can’s in more speeches. It also shows that Steve Jobs was repeating the functions in a very high frequency. He was very motivated and proud to introduce and sell the new products. He was

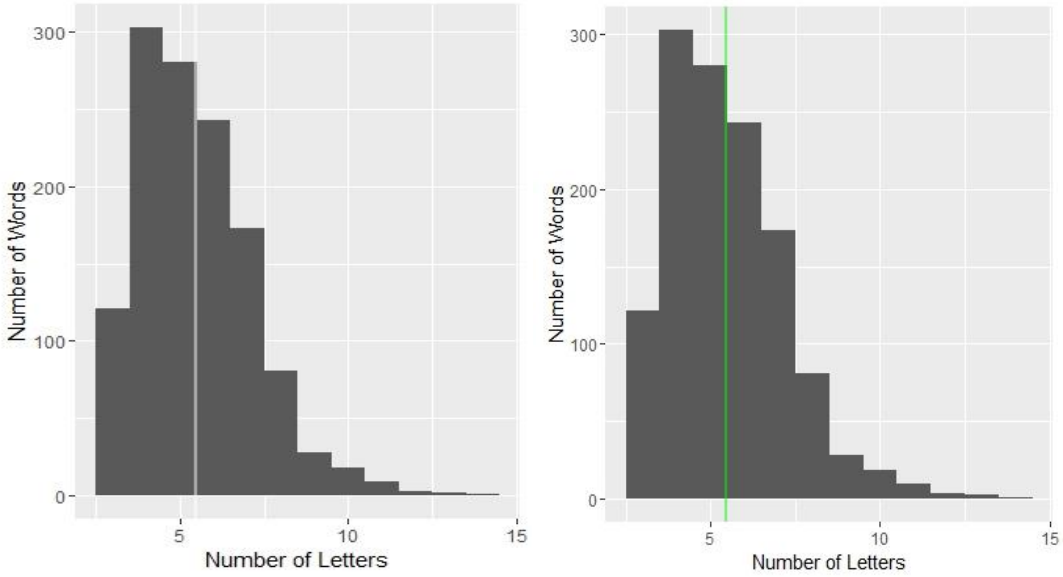
always trying to explain what his new products can do, what functions are new, what features are unique, and so on. It caused the curiosity of the audiences.

Third, in his speeches, he constantly improved the diversity of the vocabulary. From the Macintosh conference to the iPad conference, the green-small words were increasing. In other words, he was using more different words instead of only focusing on the several keywords and keeping emphasizing. In my opinion, two main reasons might cause it. First, there were more and more features. Products were waiting for him to introduce, so he had to separate his time in different words. Second, he might try to use more and more different words to explain. He is developing his speeches' artistry. And also, because his speeches in his earlier life were time limited, the first speech may not be able to represent his personal habits of speech.

Then, I find that he compared his group with IBM frequently in his first speech. I can find 'ibm' in a big size in the first speech but can't find it in the following speeches. It is easy to understand because nearly everyone knows his life is tied with the IBM. Compare with IBM is good for apple's corporate image.

Finally, I find that he likes to introduce the milestones in the timeline. We can easily find the keywords about time, such as 'now', 'year', 'day', 'today'. It is very forceful to introduce the exact event by a timeline. It also shows that Jobs did every step for a long historical ambition but quick money.

By the way, I generated the graph of 'Word Length Counts' in 1984 and 2010 because I wanted to find the conclusion of if Steve Jobs changed the word lengths significantly.



I set the mean line in 1984 in color of white and the line in 2010 in color of green. From the graphs of 'Word Length Counts', I found that most words had about 5 to 6 letters whenever in 1984 or 2010.

By summarizing the important information of these two data, I found that the medians were always 5.000 while the mean in 1984 was 5.445 and the mean in 2010 was 5.467.

Data in 1984:

```
> summary(nchar(words))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   4.000   5.000   5.445   6.000  14.000
```

Data in 2010:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   4.000   5.000   5.467   6.000  14.000
```

I also found the tables of those two charts.

Data in 1984:

```
> table(nchar(words))

  3   4   5   6   7   8   9  10  11  12  13  14
61 142 147 113  75  45  14   9   1   2   1   1
```

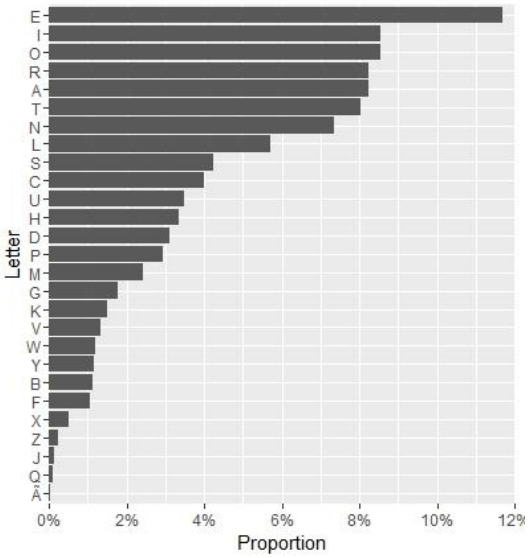
Data in 2010:

```
  3   4   5   6   7   8   9  10  11  12  13  14
121 303 280 243 173  81  28  18   9   3   2   1
```

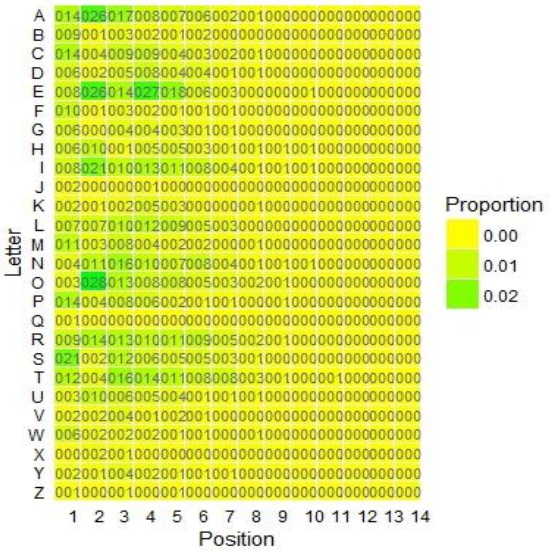
Then I got the standard deviation in two years: in 1984 is about 2.97, in 2010 is about 1.78. I set that H0: μ = μ0, H1：μ ≠ μ0, t in t-test is about 0.005. Steve Jobs uses very similar length words in the two speeches.

Finally, I also tried to build the other graphs, such as 'the frequency of letters' in 2010, and 'letter and position heatmap'.

The Frequency of Letters



Letter and Position Heatmap



I found that the least frequently used letter was 'a' and the most was 'e'.

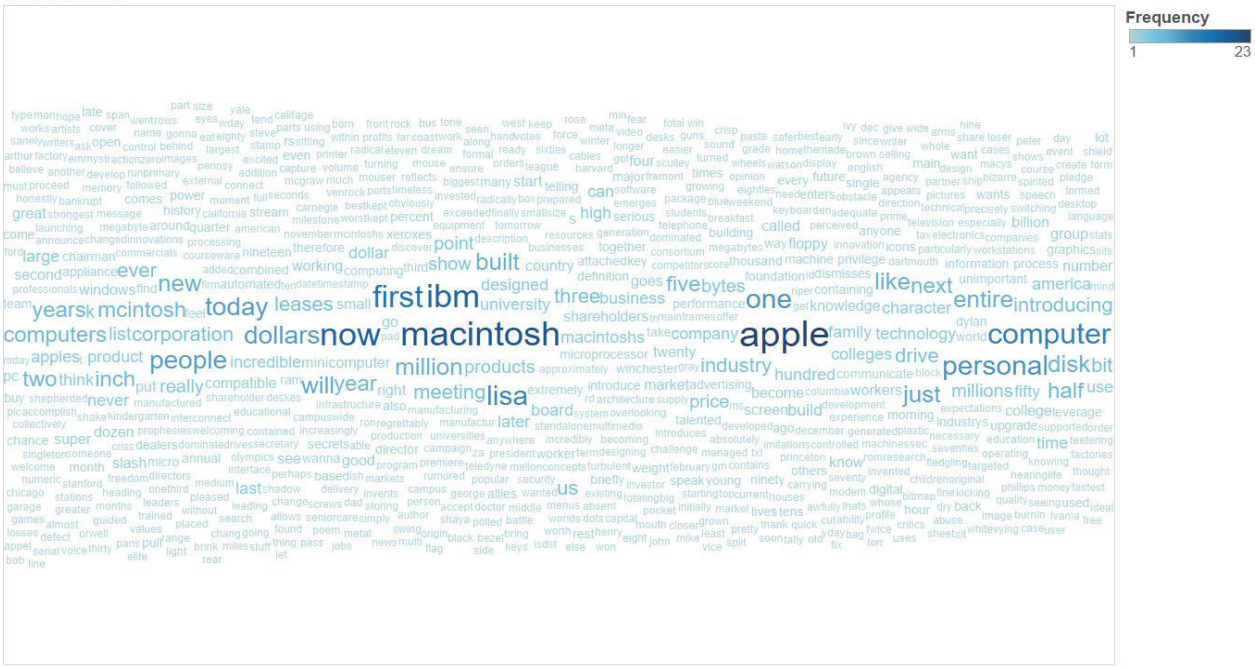
b. Tableau: Use the "cleaned" data from R. (10 points)

Prepare:

Initially, I exported all the words that had been cleaned from R. Then I downloaded the tool on this website: http://www.clearlyandsimply.com/clearly_and_simply/2015/03/the-implementation-of-word-clouds-with-excel.html . I got all the data ready with that tool.

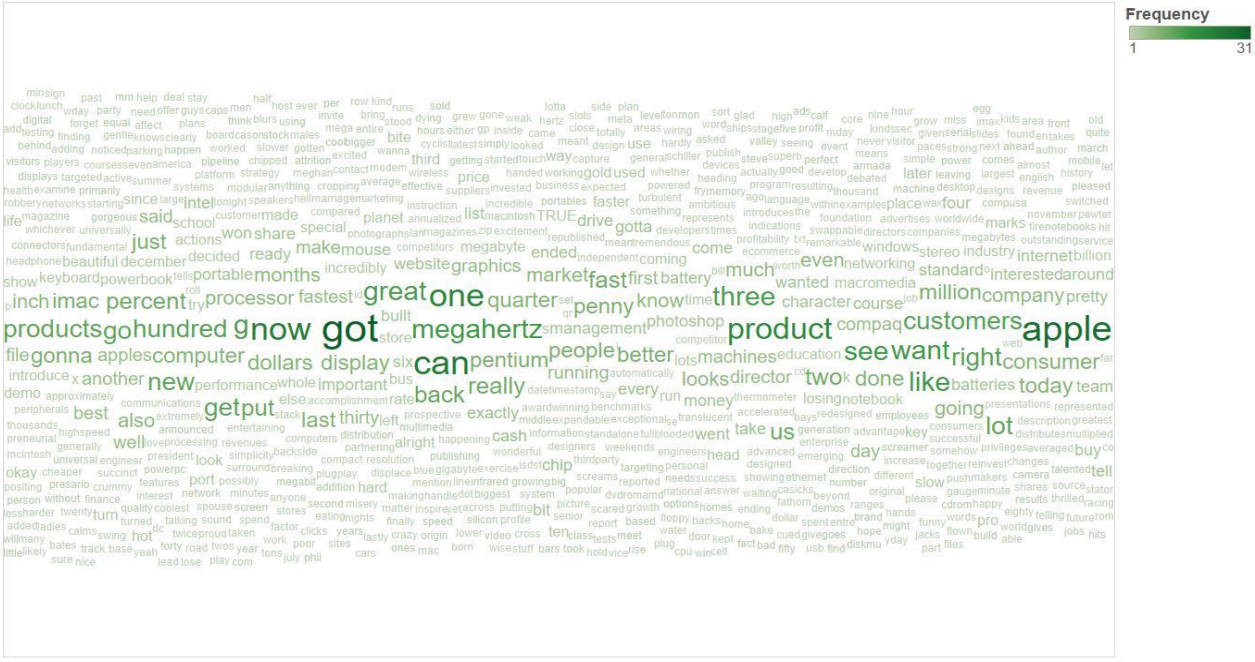
However, there was one step that I did between exporting data from R and importing data to the excel tool. I found that all the .txt files are started with “list list content c”, so I deleted all of them.

1-Macintosh



Words. Color shows sum of Frequency. Size shows sum of Frequency.

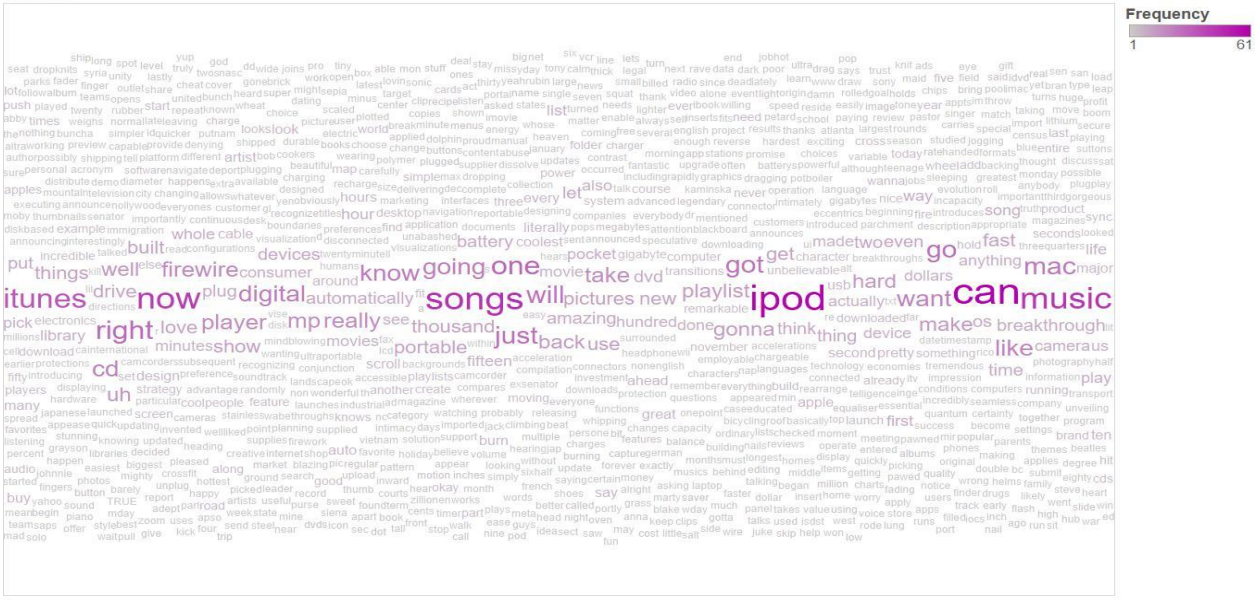
2-iMac



Words. Color shows sum of Frequency. Size shows sum of Frequency.

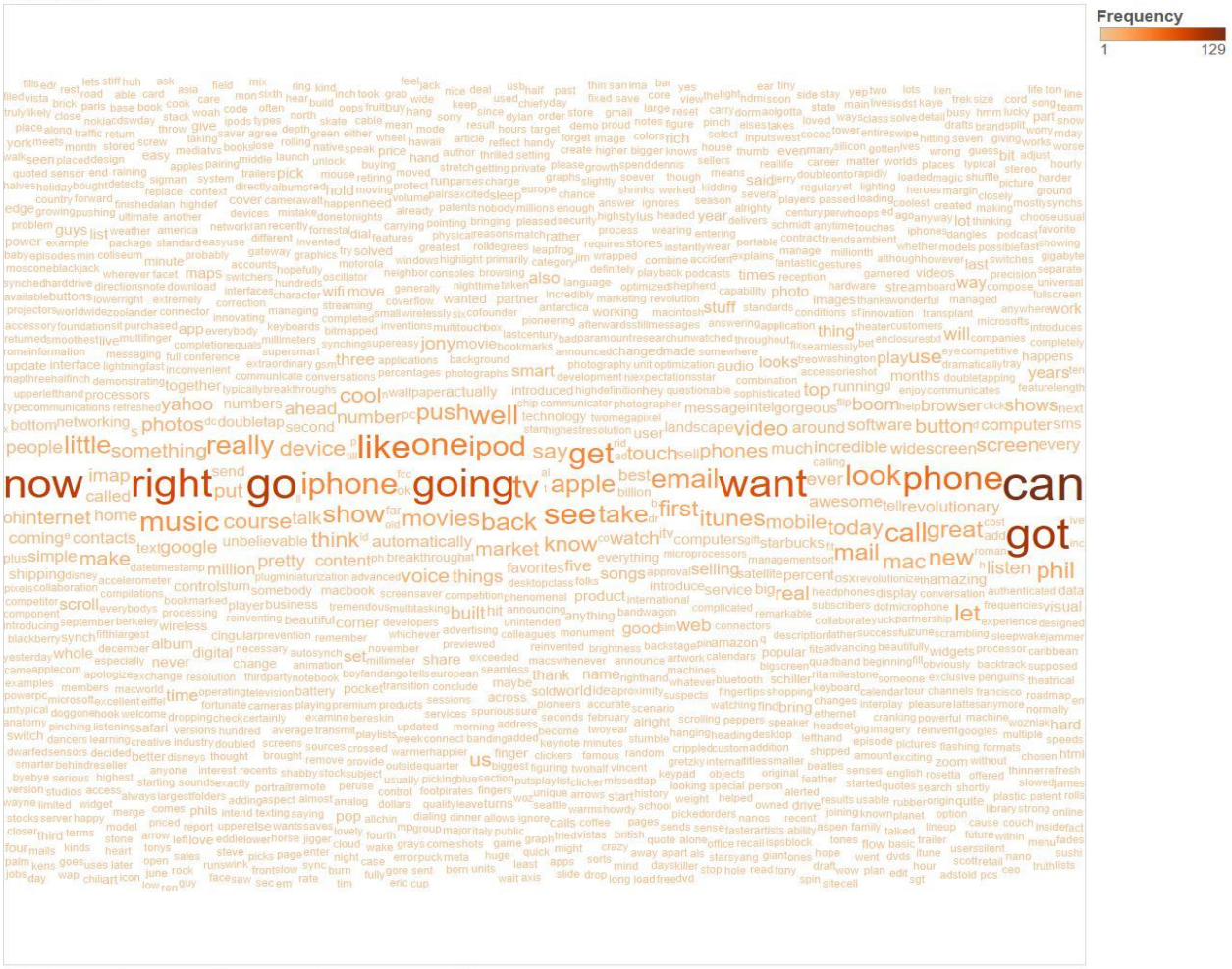
Student:Bingyang Dou Subject:CIS576 Assignment 4:Text Analysis/Visualization and Network Visualization

3-iPod



Words. Color shows sum of Frequency. Size shows sum of Frequency.

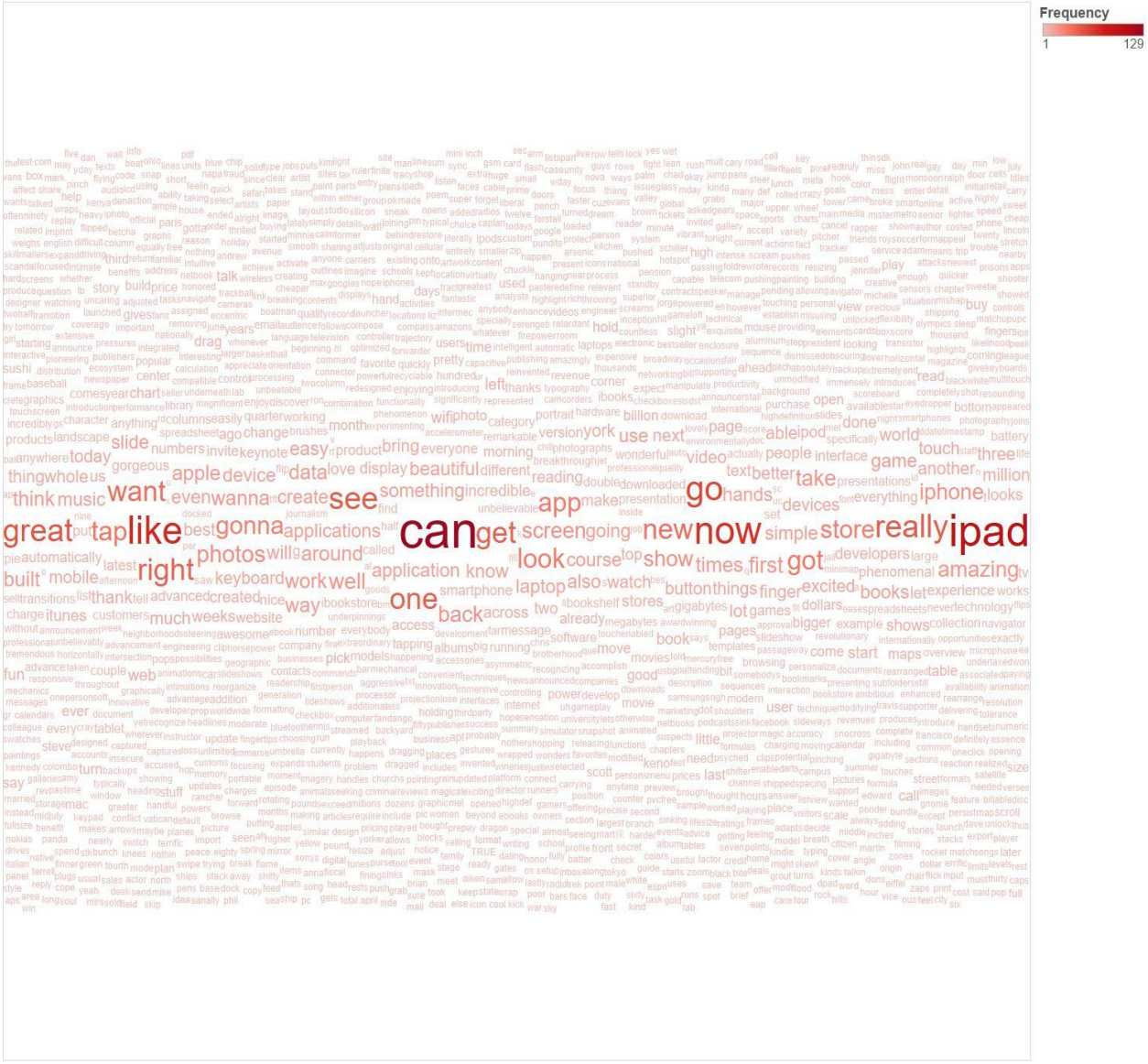
4-iPhone



Words. Color shows sum of Frequency. Size shows sum of Frequency.

Student:Bingyang Dou Subject:CIS576 Assignment 4:Text Analysis/Visualization and Network Visualization

5-iPad

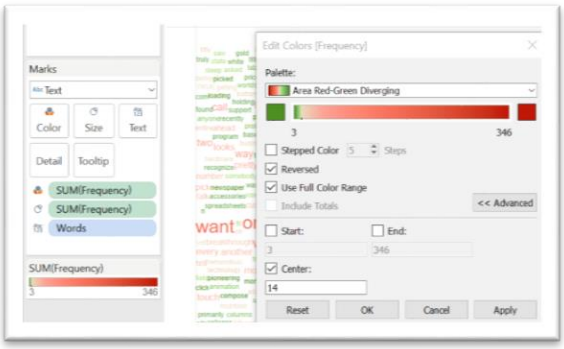


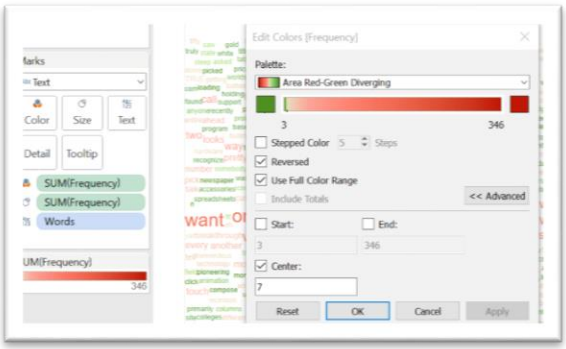
Words. Color shows sum of Frequency. Size shows sum of Frequency.

Analysis II:

The first step I did with Tableau was generating the original speeches subtitles. By comparing these word clouds, I found that, such as what I found in R's word cloud, Steve Jobs was using more and more different words in his speeches, which made his speeches more diversity. What's more, his speeches were longer and longer. Although there were some other speakers, because of the main speaker was Steve Jobs, I viewed them as the word Steve Jobs wanted to say.

To get the overall view, I also generated two word cloud graphs include all the five speeches. I merged all the five .txt files and imported the top 30% words by the reason that the whole file was too large to import. First, I set the color's Center at 14, which was the average frequency of the Top 30% words. In this setting, the words' frequency lower than 14 would be green, higher than 14 would be red. Then, I set the color's Center at 7, which was the median frequency of the Top 30% words. In this setting, the words' frequency lower than 7 would be green, higher than 7 would be red.

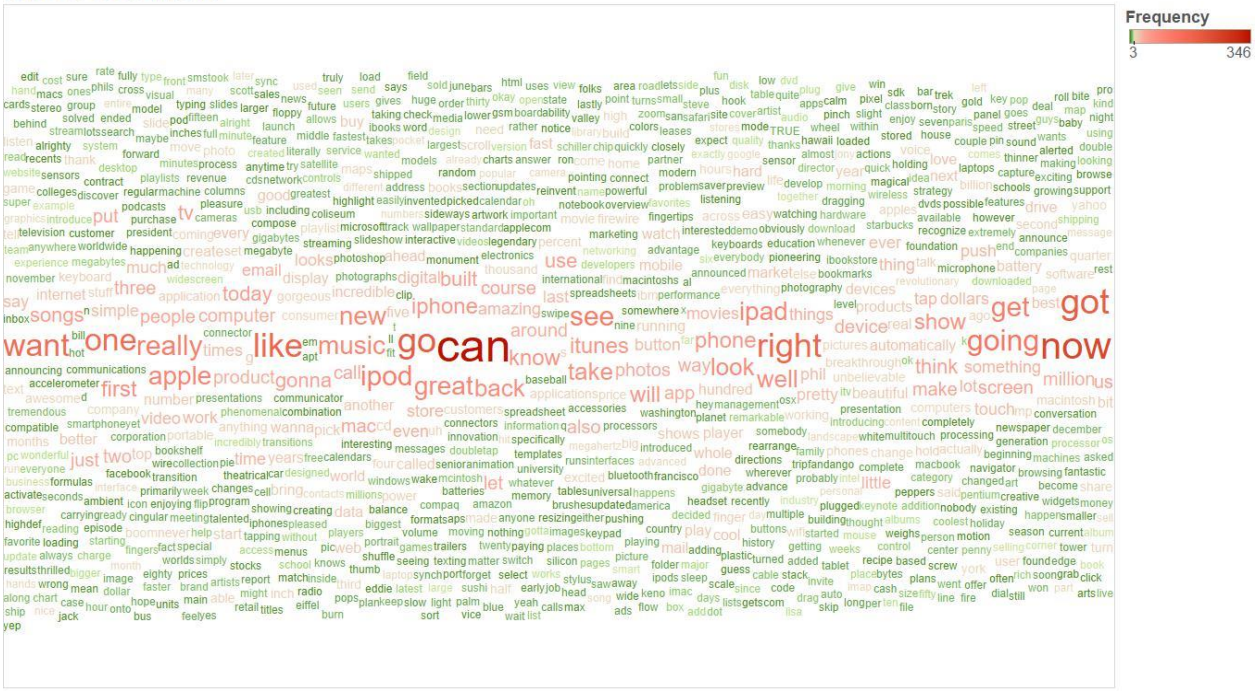




Student:Bingyang Dou Subject:CIS576 Assignment 4:Text Analysis/Visualization and Network Visualization

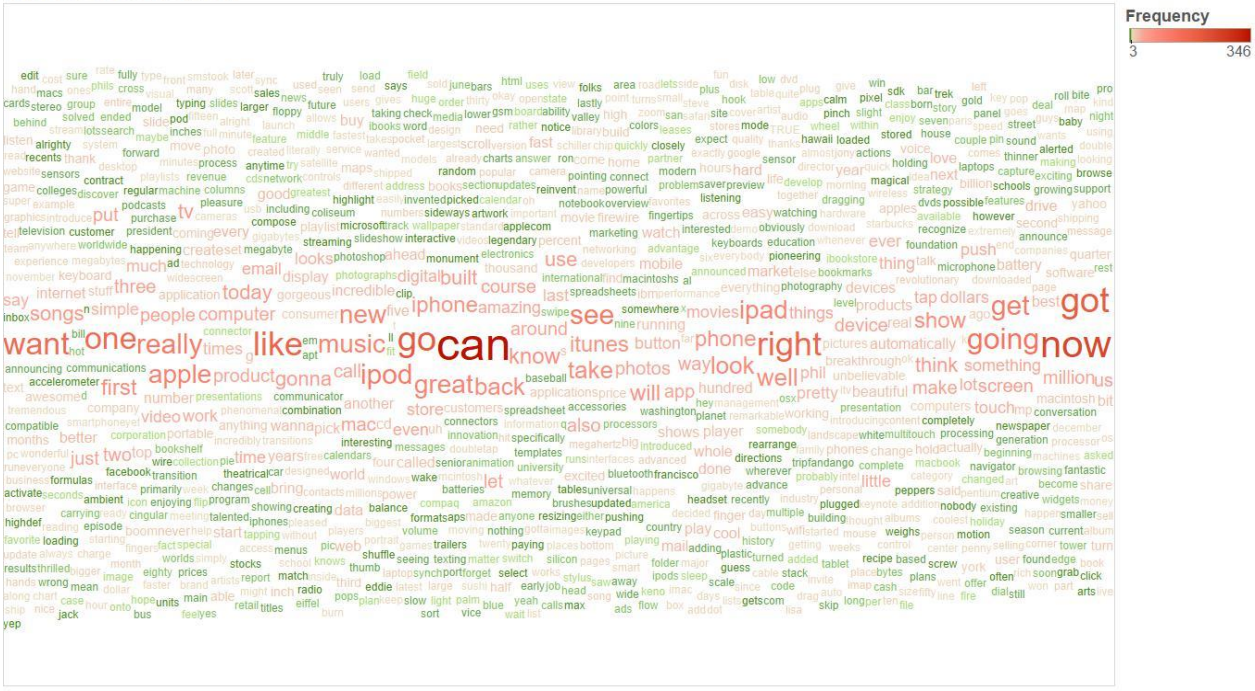
The results are shown below.

Top30%-Total-Avg14



Words. Color shows sum of Frequency. Size shows sum of Frequency.

Top30%-Total-Median7

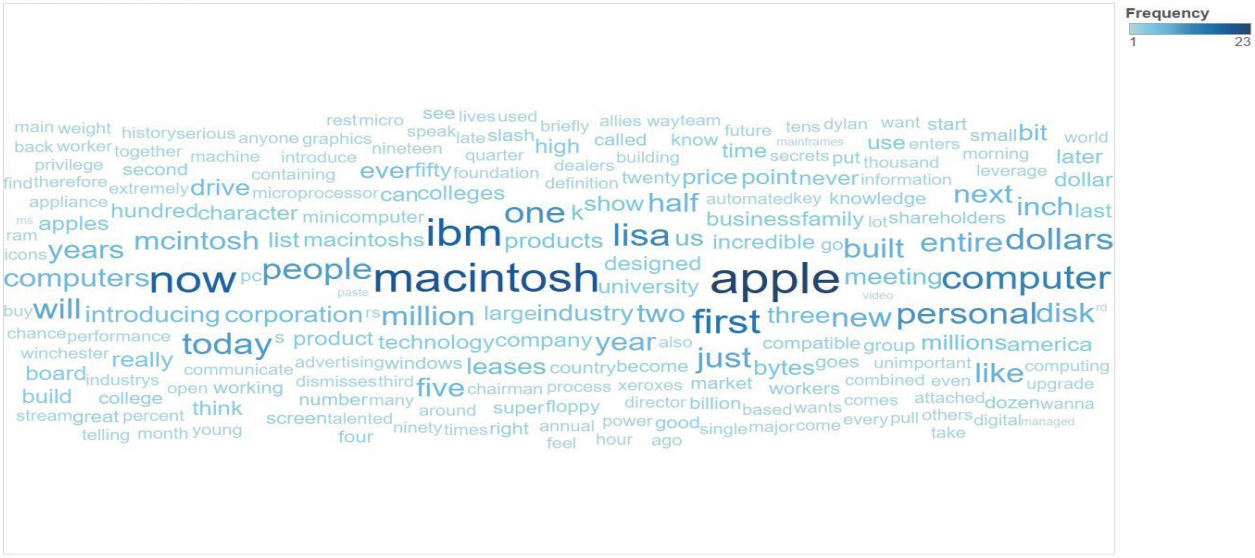


Words. Color shows sum of Frequency. Size shows sum of Frequency.

From those two graphs, I found that the one I set the center as median has more red words than green words. The median was over smaller than the average.

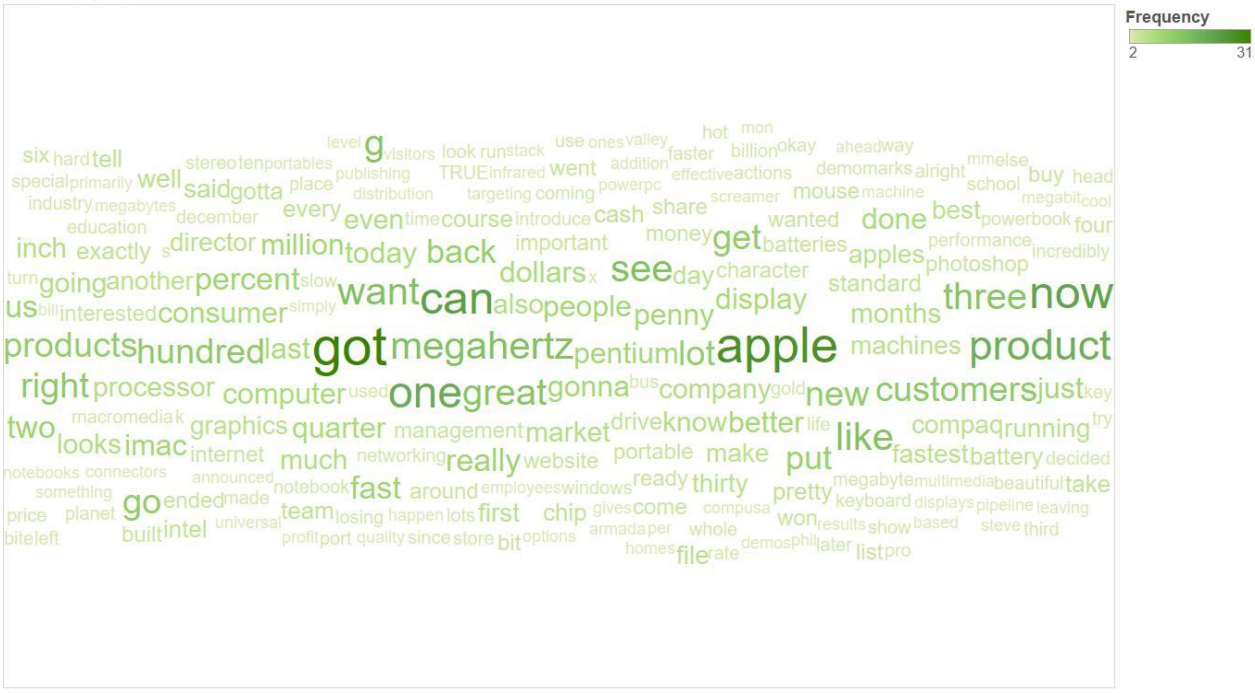
I also generated all five top 30% words' word cloud graph. Because he used around 718 words in 1984, 744 words in 1998, 1006 words in 2001, 1388 words in 2007, and 1605 words in 2010, so I imported his top 215 words in 1984, 223 words in 1998, 302 words in2001, 416 words in 2007, and 419 words in 2010. The images are shown blow.

Top30%-Macintosh



Words. Color shows sum of Frequency. Size shows sum of Frequency.

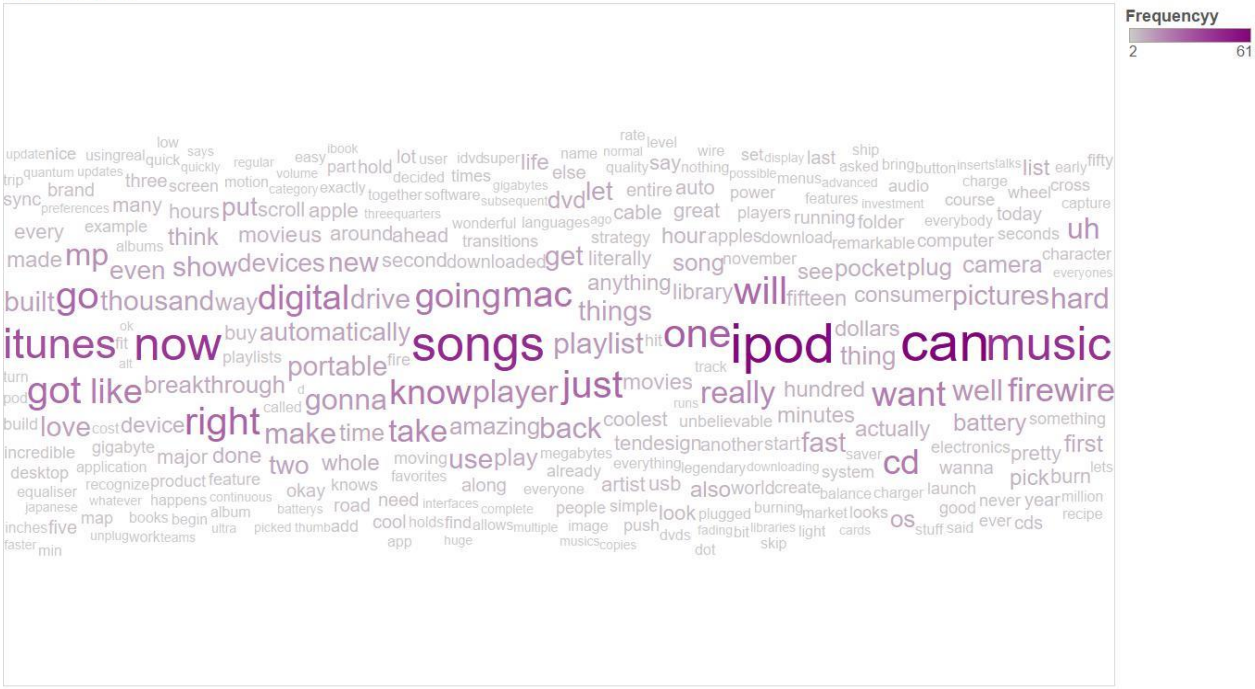
Top30%-iMac



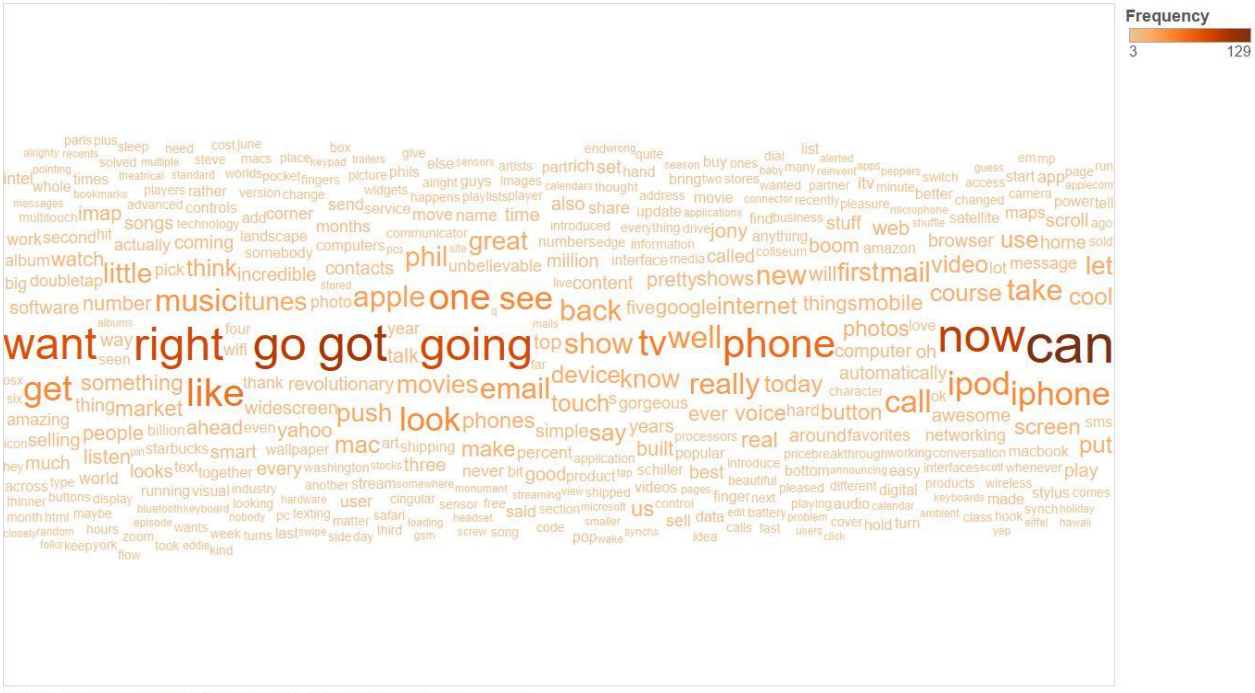
Words. Color shows sum of Frequency. Size shows sum of Frequency.

Student:Bingyang Dou Subject:CIS576 Assignment 4:Text Analysis/Visualization and Network Visualization

Top30%-iPod



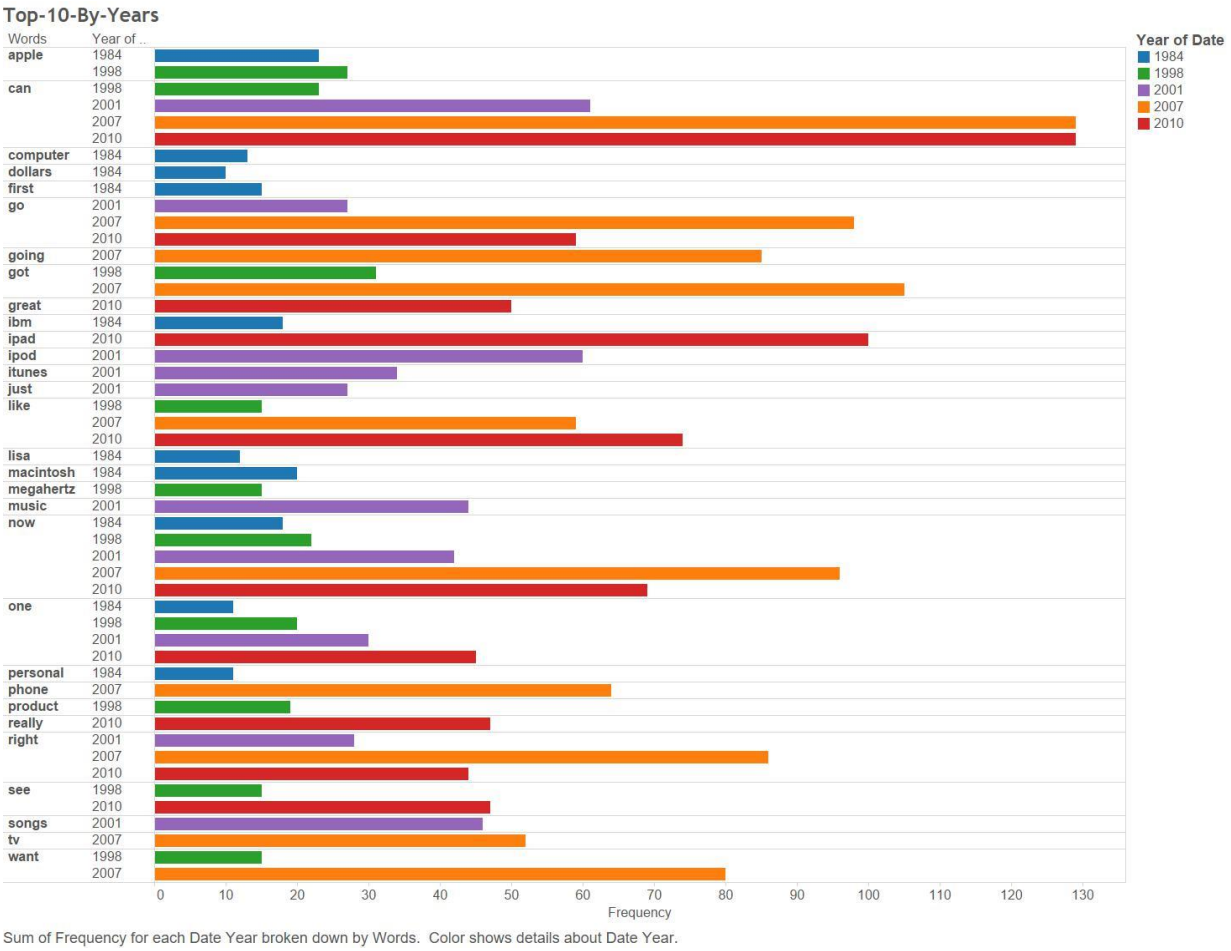
Top30%-iPhone



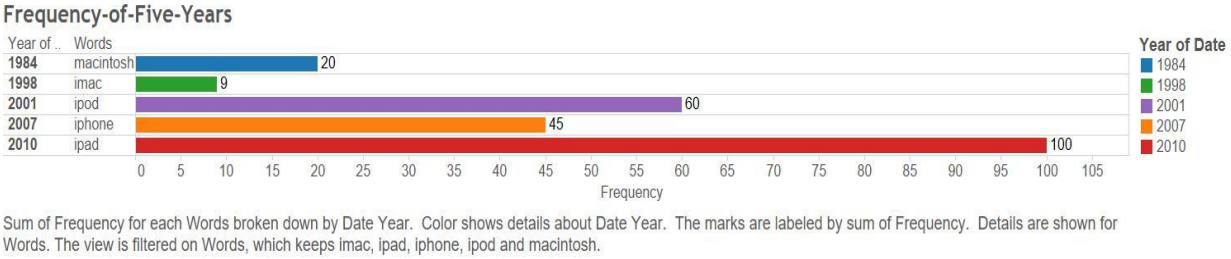
Top30%-iPad

Words. Color shows sum of Frequency. Size shows sum of Frequency.

By seeing all the images above, I can find big ‘now’ in all those five word clouds. There are also many other words relevant to time, such as ‘today’, ‘year’, in the different places. Let’s see the top ten words in each of his speech.



From the graph above, I find that there are some words only appeared more than once in the top ten words by years. They are ‘apple’, which is the brand; ‘can’, ‘like’, ‘want’, ‘right’, show attitudes; ‘now’, relevant to time; ‘go’, ‘got’, ‘see’, show some directions or achievements. Those words only appeared once are mostly used only for the certain new product in that speech.



I want to find out what percent of the products' name he used in his conference, I generated the image for the frequencies and calculated the percentage of the keywords are used: 'macintosh'(20/718=2.78%), 'imac'(9/744=1.21%), 'ipod'(60/1006=5.96%), 'iphone'(45/1388=3.24%), 'ipad'(100/1605=6.23%). The average of those percentages is 3.884%. The standard deviations are around 3.77%.

c. Tutorial:

http://georeferenced.wordpress.com/2013/01/15/rwordcloud/

d. References:

http://onertipaday.blogspot.com/2011/07/word-cloud-in-r.html
http://www.clearlyandsimply.com/clearly_and_simply/2015/03/word-clouds-withtableau.html